Duality in Bayesian prediction and its implication (Asymptotic Expansions for Various Models and Their Related Topics)

AUTHOR(S):

大西, 俊郎; 柳本, 武美

CITATION:

大西, 俊郎 ...[et al]. Duality in Bayesian prediction and its implication (Asymptotic Expansions for Various Models and Their Related Topics). 数理解析研究所講究録 2013, 1860: 104-119

ISSUE DATE:

2013-11

URL:

http://hdl.handle.net/2433/195291

数理解析研究所講究録
第1860巻 2013年 104-119

Duality in Bayesian prediction and its implication

九州大学・経済学研究院　大西俊郎
中央大学・経営システム　柳本武美
Toshio Ohnishi[a)] and Takemi Yanagimoto[b)]
[a)] Faculty of Economics, Kyushu University
[b)] Department of Industrial and Systems Engineering, Chuo University

§1. はじめに

この論文で議論するのは，実際の Bayes 分析でしばしば遭遇する次のような状況である．(1) Bayes モデル $p(x;\theta)\pi(\theta;c,\delta)$ において，超事前分布 $\lambda(c,\delta)$ を仮定する．(2) モデル $p(x;\theta,\tau)$ において，まず興味あるパラメータ θ に事前分布 $\pi(\theta|\tau)$ を仮定し，次に攪乱パラメータ τ に事前分布 $\lambda(\tau)$ を仮定する．

これらのケースを統一的に取り扱うために，$\xi\in\Xi$ というパラメータをインデックスとしてもつ Bayes モデル $p_\xi(x;\theta)\pi_\xi(\theta)$ を考え，密度 $\lambda(\xi)$ によって平均化すると考える．いわゆる **Bayesian model averaging** の枠組みである (Hoeting *et al.*, 1999)．この $\lambda(\xi)$ を **prior averaging density** と呼ぶことにする．これはモデル ξ に対する事前の信念の度合いを表すと解釈できる．

モデル ξ における周辺密度は $m_\xi(x)=\mathrm{E}[p_\xi(x;\theta)|\pi_\xi(\theta)]$ であり，全体の周辺密度は $m(x)=\mathrm{E}[m_\xi(x)|\lambda(\xi)]$ である．ただし，$\mathrm{E}[\,f(y)\,|\,p(y)\,]$ は密度 $p(y)$ の下での $f(y)$ の期

待値を表す．x をデータとする．Bayes の定理により，モデル ξ に対する事後の信念は

$$\lambda(\xi|x) = \frac{\lambda(\xi)m_\xi(x)}{m(x)} \tag{1.1}$$

のようになる．これを **posterior averaging density** と呼ぶことにする．

上記の問題を Bayes 予測問題としてとらえる．Bayes 予測問題とは，Bayes モデル $p(x;\theta)\pi(\theta)$ において，将来の確率変数 y に対する確率密度 $p(y;\theta)$ を予測分布 $q(y|x)$ によって推定することである．推定問題では予測分布が $q(y|x) = p(y;\hat{\theta})$ の形に限定されるが，予測問題では予測分布の関数形が限定されない．その意味で予測問題は推定問題の一般化と言える．

予測のよさを評価するために損失を導入する．この論文では 2 つの損失 $\mathrm{D}\big(q(y|x), p(y;\theta)\big)$ および $\mathrm{D}\big(p(y;\theta), q(y|x)\big)$ を採用し，両者の下での Bayes リスク最小問題を対比的に議論する．ここで，$\mathrm{D}\big(p(y), q(y)\big) := \mathrm{E}\big[\log\{p(y)/q(y)\} \mid p(y)\big]$ は Kullback-Leibler divergence を表す．これらの損失は互いに双対であり，それぞれ ***e*-divergence 損失**および ***m*-divergence 損失**と呼ばれる (Amari & Nagaoka, 2000).

この論文でしばしば用いる論法を 2 つ挙げておく．1 つは Pythagras 関係を用いたリスクの比較である．Pythagrean difference を

$$\mathrm{PD}(p_1, p_2, p_3) := \mathrm{D}(p_1, p_3) - \mathrm{D}(p_1, p_2) - \mathrm{D}(p_2, p_3)$$

によって定義する．もし $\mathrm{E}\big[\mathrm{PD}(p_1, p_2, p_3)\big] = 0$ ならば，不等式 $\mathrm{E}\big[\mathrm{D}(p_1, p_2)\big] \leq \mathrm{E}\big[\mathrm{D}(p_1, p_3)\big]$ が導かれる．

もう 1 つも等式を用いたリスク比較である．例を挙げる．指数型分布族 $p(x;\theta)$ において最尤推定量（MLE）$\hat{\theta}_M$ は

$$\log \frac{p(x;\hat{\theta}_M)}{p(x;\theta)} = \mathrm{D}\big(p(y;\hat{\theta}_M), p(y;\theta)\big) \tag{1.2}$$

を満たす (Kullback, 1959)．特に，$p(x;\theta)$ が多変量正規分布（3 次元以上，共分散行列は単位行列）のとき，Stein 推定量 $\hat{\theta}_S$ (Stein, 1981) は

$$\mathrm{E}\left[\log \frac{p(x;\hat{\theta}_S)}{p(x;\theta)} - \mathrm{D}\big(p(y;\hat{\theta}_S), p(y;\theta)\big) \,\middle|\, p(x;\theta)\right] = 0 \tag{1.3}$$

を満たす．これらの等式から，リスクの比較が対数尤度比の期待値の比較に帰着されることが分かる．MLE の定義により，MLE に対する対数尤度比は最大である．したがって，Stein 推定量が MLE を改善することが分かる．

(1.2) および (1.3) は対数尤度比という望大項と e-divergence 損失という望小項のバランスを意味する．このように望大項と望小項のバランスを表す等式を本論文では**鞍点等式**と呼ぶことにする．

§2. e-divergence 損失の場合

この節では e-divergence 損失を採用した場合を論じる．すなわち，Bayes 予測問題

$$\min_{q(y|x)} \mathrm{E}\Big[\mathrm{D}\big(q(y|x),\, p_\xi(y;\theta)\big) \;\Big|\; \pi_\xi(\theta|x)\lambda(\xi|x)\Big] \tag{2.1}$$

を考察する．ここで，$\pi_\xi(\theta|x)$ はモデル ξ における事後密度であり，$\lambda(\xi|x)$ は (1.1) の posterior averaging density である．$p_\xi(x;\theta)\pi_\xi(\theta)\lambda(\xi) = \pi_\xi(\theta|x)\lambda(\xi|x)m(x)$ に注意すると，$\pi_\xi(\theta|x)\lambda(\xi|x)$ は x が与えられたという条件の下での (θ,ξ) の条件付き密度であることが分かる．

Corcuera & Giummole (1999) によると，モデル ξ における Bayes 予測問題

$$\min_{q(y|x)} \mathrm{E}\Big[\mathrm{D}\big(q(y|x),\, p_\xi(y;\theta)\big) \;\Big|\; \pi_\xi(\theta|x)\Big]$$

の解は

$$q^e_\xi(y|x) \propto \exp\big\{\mathrm{E}\big[\log p_\xi(y;\theta) \;\big|\; \pi_\xi(\theta|x)\big]\big\} \tag{2.2}$$

で与えられる．最適性は Pythagras 関係

$$\mathrm{E}\Big[\mathrm{PD}\big(q(y|x),\, q^e_\xi(y|x),\, p_\xi(y;\theta)\big) \;\Big|\; \pi_\xi(\theta|x)\Big] = 0 \tag{2.3}$$

によって明瞭に理解できる (Yanagimoto & Ohnishi, 2009).

Bayes 予測問題 (2.1) を少し一般化して，リスク最小問題

$$\min_{q(y|x)} \mathrm{E}\Big[\mathrm{D}\big(q(y|x),\, q^e_\xi(y|x)\big) \;\Big|\; h(\xi)\Big] \tag{2.4}$$

を考える．ただし，$h(\xi)$ は適当な密度であり，**canonical weight** と呼ぶことにする．リスク最小問題 (2.4) は次のようにして得られる．Bayes 予測問題 (2.1) は (2.3) を用いると，

$$\min_{q(y|x)} \mathrm{E}\Big[\mathrm{D}\big(q(y|x),\, q_\xi^e(y|x)\big) \;\Big|\; \lambda(\xi|x)\Big]$$

のように等価変形できる．ここで，posterior averaging density $\lambda(\xi|x)$ を canonical weight $h(\xi)$ に置き換えればよい．

以下の議論で重要な役割を果たす量を定義しておく．

Definition 2.1. (i) (2.2) の $q_\xi^e(y|x)$ を用いて

$$f^e(y|x;h) := \exp\big\{\mathrm{E}\big[\log q_\xi^e(y|x) \;\big|\; h(\xi)\big] - \psi_x(h)\big\} \tag{2.5}$$

とおく．ここで，$\exp\{\psi_x(h)\}$ は規格化因子である．予測分布 $f^e(y|x;h)$ を canonical weight h による $q_\xi^e(y|x)$ の ***e*-mixture** と呼ぶ．

(ii) 次の量を canonical weight h に対応する **mean weight** と呼ぶ．

$$t_x(\xi;h) := \mathrm{E}\big[\log q_\xi^e(y|x) \;\big|\; f^e(y|x;h)\big]. \tag{2.6}$$

$f^e(y|x;h)$，$\psi_x(h)$ および $t_x(\xi;h)$ は h の汎関数である．

$\log q(x|x)$ は対数尤度の拡張と考えるので，**Bayesian log-likelihood** と呼ぶ．Bayesian log-likelihood ratio と e-divergence 損失のバランスを意味する等式が得られる．

Theorem 2.1. (2.5) の $f^e(y|x;h)$ は Pythagras 関係

$$\mathrm{E}\big[\mathrm{PD}\big(q(y|x),\, f^e(y|x;h),\, q_\xi^e(y|x)\big) \;\big|\; h(\xi)\big] = 0$$

を満たす．したがって，リスク最小問題 (2.4) の解である．また，鞍点等式

$$\mathrm{E}\left[\log \frac{f^e(x|x;h)}{q_\xi^e(x|x)} - \mathrm{D}\big(f^e(y|x;h),\, q_\xi^e(y|x)\big) \;\middle|\; h(\xi)\right] = 0 \tag{2.7}$$

が成り立つ.

(2.7) からリスク最小問題 (2.4) の最小値が $-\psi_x(h)$ であることが分かる.

以下でしばしば用いる Gateaux 微分の記法を定めておく. $F(h)$ を canonical weight h の汎関数とする. $F(h)$ の h_1 における増分 h_2-h_1 に対する Gateaux 微分を $\delta_G F(h_1; h_2-h_1)$ と書く. すなわち,

$$\delta_G F(h_1; h_2-h_1) := \lim_{\beta\to 0} \frac{F\big(h_1+\beta(h_2-h_1)\big)-F(h_1)}{\beta}$$

である. Canonical weight h が離散型の場合, Gateaux 微分は普通の偏微分になり, 以下の議論は Ohnishi & Yanagimoto (2013) に帰着される. Mean weight (2.6) は $\psi_x(h)$ の Gateaux 微分に現れる.

$$\delta_G \psi_x(h_1; h_2-h_1) = \mathrm{E}\big[t_x(\xi; h_1) \,\big|\, h_2(\xi)-h_1(\xi)\big].$$

次の定理は, 最適予測分布を求めることがある制約条件の下で Shannon entropy を最大化することと等価であることを示している. $\mathrm{H}\big[p(y)\big] := \mathrm{E}\big[-\log p(y) \,\big|\, p(y)\big]$ は確率密度 $p(y)$ の Shannon entropy とする.

Theorem 2.2. $s(\xi) = t_x(\xi; h)$ のときに限り, Shannon entropy の制約付き最大問題

$$\begin{aligned} &\max \mathrm{H}\big[q(y|x)\big] \\ &\text{subject to } \mathrm{E}\big[\log q_\xi^e(y|x) \,\big|\, q(y|x)\big] = s(\xi) \end{aligned}$$

は, リスク最小問題 (2.4) と同一の解 $f^e(y|x; h)$ をもつ.

証明の本質を説明する. $p_1(y), p_2(y)$ を所与の確率密度とする. 指数型分布族

$$p(y;\eta) = \exp\left\{\eta \log\frac{p_1(y)}{p_2(y)} - \psi(\eta)\right\} p_2(y)$$

は, 次の 2 つの問題の解として得られる. ただし, $\mu = \psi'(\eta)$ とする.

1) Shannon entropy の制約付き最大問題

$$\max \mathrm{E}\left[-\frac{q(y)}{p_2(y)}\log\frac{q(y)}{p_2(y)}\;\middle|\;p_2(y)\right],$$
$$\text{subject to } \mathrm{E}\left[\log\frac{p_1(y)}{p_2(y)}\;\middle|\;q(y)\right]=\mu.$$

2) e-divergence の線形結合の最小問題

$$\min_{q(y|x)}\Big\{\eta\mathrm{D}\big(q(y),p_1(y)\big)+(1-\eta)\mathrm{D}\big(q(y),p_2(y)\big)\Big\}.$$

Bayesian log-likelihood を‘最大化’することよって面白い予測分布が得られる.

Theorem 2.3. Canonical weight $h_x^\dagger(\xi)$ を次によって定義する.

$$\delta_G \log f^e(x|x;h_x^\dagger;h-h_x^\dagger)=0 \quad \text{for any } h. \tag{2.8}$$

このとき，予測分布 $f^e(y|x;h_x^\dagger)$ は次の鞍点等式を満たす.

$$\log\frac{f^e(x|x;h_x^\dagger)}{q_\xi^e(x|x)}=\mathrm{D}\big(f^e(y|x;h_x^\dagger),\,q_\xi^e(y|x)\big) \quad \text{for any } \xi.$$

$h_x^*(\xi)$ を (1.1) の posterior averaging density とする．すなわち，$h_x^*(\xi):=\lambda(\xi|x)$ とおく．Theorem 2.1 から $f^e(y|x;h_x^*)$ は Bayes 予測問題 (2.1) の解，すなわち，最適予測分布である．これを‘トップ’として含み，Theorem 2.3 の予測分布を‘ビリ’として含むような予測分布のクラスが鞍点等式を通じて規定される.

Theorem 2.4. 次の鞍点等式を満たす予測分布のクラスを $\mathcal{Q}^e$ と書く.

$$\mathrm{E}\left[\log\frac{f^e(x|x;h)}{q_\xi^e(x|x)}-\mathrm{D}\big(f^e(y|x;h),\,q_\xi^e(y|x)\big)\;\middle|\;\lambda(\xi|x)m(x)\right]=0. \tag{2.9}$$

また，(2.8) の $h_x^\dagger(\xi)$ が実際に Bayesian log-likelihood $\log f^e(x|x;h)$ を最大化するとする．このとき，$\mathcal{Q}^e$ の中で，$f^e\big(y|x;h_x^*\big)$ は最良であり，$f^e\big(y|x;h_x^\dagger\big)$ は最悪である.

平たく言えば，Bayesian log-likelihood の最大化が予測の最低保証を与えると解釈できる．Theorem 2.3 の予測分布は，指数型分布族における MLE と同じような役割を果た

す．Yanagimoto & Ohnishi (2011) は条件 (2.9) に着目し，情報量基準との関連を論じている．

リスク最小問題 (2.4) の最小値 $-\psi_x(h)$ を h について‘最大化’することにより，事後リスクを一定にするような予測分布を導くことができる．

Theorem 2.5. Canonical weight $h_x^{\mathrm{c}}(\xi)$ を次によって定義する．

$$\delta_G\psi_x(h_x^{\mathrm{c}}; h - h_x^{\mathrm{c}}) = 0 \quad \text{for any } h.$$

このとき，予測分布 $f^e(y|x; h_x^{\mathrm{c}})$ は次の等式を満たす．

$$\mathrm{D}\big(f^e(y|x; h_x^{\mathrm{c}}),\, q_\xi^e(y|x)\big) = -\psi_x(h_x^{\mathrm{c}}).$$

予測分布 $f^e(y|x; h_x^{\mathrm{c}})$ はどの $q_\xi^e(y|x)$ からも‘等距離’なので，任意の prior averaging density に対して事後リスクが一定値 $-\psi_x(h_x^{\mathrm{c}})$ をとることになる．

§3. m-divergence 損失の場合

この節では e-divergence 損失と双対な m-divergence 損失を採用し，Bayes 予測問題

$$\min_{q(y|x)} \mathrm{E}\Big[\mathrm{D}\big(p_\xi(y;\theta),\, q(y|x)\big) \;\Big|\; \pi_\xi(\theta|x)\lambda(\xi|x)\Big] \tag{3.1}$$

を考察する．Shannon entropy の差と m-divergence のバランスを意味する等式が重要な役割を果たす．この節の議論と §2 が対数尤度最大化と Shannon entropy 最大化の間に内在する双対性を浮き彫りにする．

Aitchson (1975) によれば，モデル ξ における Bayes 予測問題

$$\min_{q(y|x)} \mathrm{E}\Big[\mathrm{D}\big(p_\xi(y;\theta),\, q(y|x)\big) \;\Big|\; \pi_\xi(\theta|x)\Big]$$

の解は

$$q_\xi^m(y|x) := \mathrm{E}\big[p_\xi(y;\theta) \;\big|\; \pi_\xi(\theta|x)\big] \tag{3.2}$$

で与えらえる．また，最適性は Pythagras 関係

$$\mathrm{E}\Big[\mathrm{PD}\big(p_\xi(y;\theta),\, q_\xi^m(y|x),\, q(y|x)\big) \;\Big|\; \pi_\xi(\theta|x)\Big] = 0 \tag{3.3}$$

を通じて明瞭に示される (Yanagimoto & Ohnishi, 2009).

§2 と同様に問題を一般化し，

$$\min_{q(y|x)} \mathrm{E}\Big[\mathrm{D}\big(q_\xi^m(y|x),\, q(y|x)\big) \;\Big|\; h(\xi)\Big] \tag{3.4}$$

を考える．ここでも $h(\xi)$ を **canonical weight** と呼ぶ．リスク最小問題 (3.4) は，(3.3) を用いて (3.1) を

$$\min_{q(y|x)} \mathrm{E}\Big[\mathrm{D}\big(q_\xi^m(y|x),\, q(y|x)\big) \;\Big|\; \lambda(\xi|x)\Big]$$

のように等価変形し，$\lambda(\xi|x)$ を $h(\xi)$ に置き換えば得られる．

以下の議論で重要な役割を果たす量を定義する．

Definition 3.1. (i) (3.2) の $q_\xi^m(y|x)$ を用いて

$$f^m(y|x;h) := \mathrm{E}\big[q_\xi^m(y|x) \;\big|\; h(\xi)\big] \tag{3.5}$$

とおく．予測分布 $f^m(y|x;h)$ を canonical weight h による $q_\xi^m(y|x)$ の m-**mixture** と呼ぶ．

(ii) 次の量を canonical weight h に対応する **entropy weight** と呼ぶ．

$$t_x(\xi;h) = -\log f^m(x|x;h) - \mathrm{D}\big(q_\xi^m(y|x),\, f^m(y|x;h)\big). \tag{3.6}$$

リスク最小問題 (3.4) の解は Shannon entropy の差と m-divergence 損失をバランスさせる．

Theorem 3.1. (i) (3.5) の $f^m(y|x;h)$ は Pythagras 関係

$$\mathrm{E}\big[\mathrm{PD}\big(q_\xi^m(y|x),\, f^m(y|x;h),\, q(y|x)\big) \;\big|\; h(\xi)\big] = 0 \tag{3.7}$$

を満たす．したがって，リスク最小問題 (3.4) の解である．また，$f^m(y|x;h)$ は次の鞍点等式を満たす．

$$\mathrm{E}\Big[\mathrm{H}\big[f^m(y|x;h)\big] - \mathrm{H}\big[q_\xi^m(y|x)\big] - \mathrm{D}\big(q_\xi^m(y|x),\, f^m(y|x;h)\big) \;\Big|\; h(\xi)\Big] = 0. \tag{3.8}$$

汎関数 $\psi_x(h)$ を

$$-\psi_x(h) := \mathrm{H}\big[f^m(y|x;h)\big] - \mathrm{E}\Big[\mathrm{H}\big[q_\xi^m(x|x)\big] \;\Big|\; h(\xi)\Big]$$

によって定義する．(3.8) から，$-\psi_x(h)$ はリスク最小問題 (3.4) の最小値であることが分かる．興味深いことに，$\psi_x(h)$ の Gateaux 微分は entropy weight を用いて表すことができる．

$$\delta_G\psi_x(h_1; h_2 - h_1) = \mathrm{E}\big[t_x(\xi;h_1) \;\big|\; h_2(\xi) - h_1(\xi)\big].$$

リスク最小問題 (3.4) を等価な最大問題に書き換える．ここで等価とは解が同一という意味である．

Theorem 3.2. $s(\xi) = t_x(\xi;h)$ のときに限り，Bayesian log-likelihood の制約付き最大問題

$$\begin{aligned} &\max_{q(y|x)} \ \log q(x|x) \\ &\text{subject to } -\log q(x|x) - \mathrm{D}\big(q_\xi^m(y|x),\, q(y|x)\big) = s(\xi) \end{aligned}$$

はリスク最小問題 (3.4) と同一の解 $f^m(y|x;h)$ をもつ．

Shannon entropy を‘最大化’すると (3.8) とは別の鞍点等式が得られる．

Theorem 3.3. Canonical weight $h_x^\dagger(\xi)$ を次によって定義する．

$$\delta_G\mathrm{H}\big[f^m(y|x;h_x^\dagger;h-h_x^\dagger)\big] = 0 \quad \text{for any } h. \tag{3.9}$$

このとき，予測分布 $f^m(y|x;h_x^\dagger)$ は次の鞍点等式を満たす．

$$\mathrm{H}\big[f^m(y|x;h_x^\dagger)\big] - \mathrm{H}\big[q_\xi^m(y|x)\big] = \mathrm{D}\big(q_\xi^m(y|x),\, f^m(y|x;h_x^\dagger)\big) \quad \text{for any } \xi.$$

§2 と同様に，$h_x^*(\xi) = \lambda(\xi|x)$ とおく．

Theorem 3.4. 次の鞍点等式を満たす予測分布のクラスを $\mathcal{Q}^m$ と書く．

$$\mathrm{E}\Big[\mathrm{H}\big[f^m(y|x;h)\big] - \mathrm{H}\big[q_\xi^m(y|x)\big] - \mathrm{D}\big(q_\xi^m(y|x),\, f^m(y|x;h)\big) \;\Big|\; \lambda(\xi|x)m(x)\Big] = 0.$$

また，(3.9) の $h_x^\dagger(\xi)$ が実際に Shannon entropy $\mathrm{H}\big[f^m(y|x;h)\big]$ を最大化するとする．このとき，$\mathcal{Q}^m$ の中で，$f^m(y|x;h_x^*)$ は最良であり，$f^m(y|x;h_x^\dagger)$ は最悪である．

リスク最小問題 (3.4) の最小値 $-\psi_x(h)$ を h について‘最大化’することにより，事後リスクが一定という意味で頑健な予測分布を導くことができる．

Theorem 3.5. Canonical weight $h_x^{\mathrm{c}}(\xi)$ を次によって定義する．

$$\delta_G\psi_x(h_x^{\mathrm{c}};\, h - h_x^{\mathrm{c}}) = 0 \quad \text{for any } h.$$

このとき，予測分布 $f^m(y|x;h_x^{\mathrm{c}})$ は次の等式を満たす．

$$\mathrm{D}\big(q_\xi^m(y|x),\, f^m(y|x;h_x^{\mathrm{c}})\big) = -\psi_x(h_x^{\mathrm{c}}) \quad \text{for any } \xi.$$

§4. 双対性の含意

§4.1. Mean weight と entropy weight

Definition 2.1 (ii) の (2.6) で定義した mean weight $t_x(\xi;h)$ について考察する．この量は

$$t_x(\xi;h) = -\mathrm{H}\big[f^e(y|x;h)\big] - \mathrm{D}\big(f^e(y|x;h),\, q_\xi^e(y|x)\big)$$

のように表現することができるので，モデル ξ に対する選好度を表していると解釈できる．

(2.8) によって定義される $h_x^\dagger$ を陽に書き下すことはできないが，対応する mean weight は簡単な表現をもつ．Theorem 2.3 の系として次を得る．

Corollary 4.1. Canonical weight $h_x^\dagger$ に対応する mean weight は

$$t_x(\xi; h_x^\dagger) = \log q_\xi^e(x|x) - A_x\big[f^e(y|x; h_x^\dagger)\big]$$

である．ただし，$A_x\big[p(y)\big] := \log p(x) + \mathrm{H}\big[p(y)\big]$ である．

この式をもう少し詳しくみるために，$p_\xi(x;\theta)$ が指数型分布族の場合を考える．モデル ξ における MLE を $\hat\theta_{M\xi}$ とすると，

$$\log q_\xi^e(x|x) = \log p_\xi(x; \hat\theta_{M\xi}) - \mathrm{D}\big(p_\xi(y; \hat\theta_{M\xi}),\, q_\xi^e(y|x)\big)$$

が成り立つ．したがって，

$$t_x(\xi; h_x^\dagger) = \log p_\xi(x; \hat\theta_{M\xi}) - \big\{\mathrm{D}\big(p_\xi(y; \hat\theta_{M\xi}),\, q_\xi^e(y|x)\big) + A_x\big[f^e(y|x; h_x^\dagger)\big]\big\}$$

となる．これは AIC (Akaike, 1973) と同様に (最大対数尤度) − (罰則項) の形をしている．

次に，Definition 3.1 (ii) の (3.6) で定義した entropy weight について考察する．Entropy weight の定義式 (3.6) から，これもモデル ξ に対する選好度を表していると解釈できる．

Theorem 3.3 の系として次を得る．

Corollary 4.2. (3.9) で定義される canonical weight $h_x^\dagger$ に対応する entropy weight は

$$t_x(\xi; h_x^\dagger) = \mathrm{H}\big[q_\xi^m(y|x)\big] - A_x\big[f^m(y|x; h_x^\dagger)\big]$$

である．

Corollary 4.2 の意味をもう少し調べるために，$p_\xi(y;\theta)$ が混合モデルのときを考える．$\theta_{M\xi} = \operatorname{argmax} \mathrm{H}[\,p_\xi(y;\theta)\,]$ とおく．簡単な計算から，

$$\mathrm{H}\big[q_\xi^m(y|x)\big] = \mathrm{H}\big[p_\xi(y;\theta_{M\xi})\big] - \mathrm{D}\big(p_\xi(y;\theta_{M\xi}),\, q_\xi^m(y|x)\big)$$

であるので，

$$t_x(\xi; h_x^\dagger) = \mathrm{H}\big[p_\xi(y;\theta_{M\xi})\big] - \Big\{\mathrm{D}\big(p_\xi(y;\theta_{M\xi}),\, q_\xi^m(y|x)\big) + A_x\big[f^m(y|x;h_x^\dagger)\big]\Big\}$$

となる．興味深いことに，(最大 Shannon entropy) − (罰則項) の形をしている．

§4.2. α-divergence 損失の場合

最後に **α-divergence 損失**の場合を考える．この損失は次のように定義される．

$$\mathrm{D}_\alpha\big(p(y;\theta),\, q(y|x)\big) := \mathrm{E}\left[u_\alpha\left(\frac{q(y|x)}{p(y;\theta)}\right)\,\middle|\, p(y;\theta)\right],$$
$$u_\alpha(r) := \frac{4}{1-\alpha^2}\left(1 - r^{\frac{1+\alpha}{2}}\right).$$

ただし，$-1 < \alpha < 1$ である．α-divergence は，e-divergence および m-divergence の拡張である．$u_1(r) := r\log r$ および $u_{-1}(r) := -\log r$ のように定義すると，e-divergence は $\alpha = +1$，m-divergence は $\alpha = -1$ の場合と考えられる．

α-divergence 損失の下での Bayes 予測問題は

$$\min_{q(y|x)} \mathrm{E}\Big[\mathrm{D}_\alpha\big(p_\xi(y;\theta),\, q(y|x)\big)\,\Big|\,\pi_\xi(\theta|x)\lambda(\xi|x)\Big] \tag{4.1}$$

である．§2 および §3 と同様に Yanagimoto & Ohnishi (2009) の結果を用いると，これをリスク最小問題

$$\min_{q(y|x)} \mathrm{E}\Big[\mathrm{D}_\alpha\big(q_\xi^\alpha(y|x),\, q(y|x)\big)\,\Big|\, h(\xi)\Big] \tag{4.2}$$

に書き換えることができる．ここで，$q_\xi^\alpha(y|x)$ は，モデル ξ における Bayes 予測問題

$$\min_{q(y|x)} \mathrm{E}\Big[\mathrm{D}_\alpha\big(p_\xi(y;\theta),\, q(y|x)\big)\,\Big|\,\pi_\xi(\theta|x)\Big]$$

の解であり，

$$q_\xi^\alpha(y|x) \propto \Big(\mathrm{E}\big[\{p(y;\theta)\}^{\frac{1-\alpha}{2}}\,\big|\,\pi_\xi(\theta|x)\big]\Big)^{\frac{2}{1-\alpha}}$$

で与えられる (Corcuera & Giummole, 1999). 本小節でも (4.2) における $h(\xi)$ を **canonical weight** と呼ぶ.

Definition 4.1. (i) 次の予測分布を, canonical weight h による $q_\xi^\alpha(y|x)$ の **α-mixture** と呼ぶ.

$$f^\alpha(y|x;h) := \frac{1}{c_x(h)}\Big(\mathrm{E}\big[\{q_\xi^\alpha(y|x)\}^{\frac{1-\alpha}{2}} \mid h(\xi)\big]\Big)^{\frac{2}{1-\alpha}}. \tag{4.3}$$

ここで, $c_x(h)$ は規格化因子である.

(ii) 次の量を canonical weight h に対応する **divergence weight** と呼ぶ.

$$t_x(\xi;h) = u_\alpha\big(f^\alpha(x|x;h)\big) - \mathrm{D}_\alpha\big(q_\xi^\alpha(y|x), f^\alpha(y|x;h)\big). \tag{4.4}$$

§2 および §3 の定理に対応する定理が得られる.

Theorem 4.1. (4.3) の $f^\alpha(y|x;h)$ はリスク最小問題 (4.2) の最適解であり, 次の等式を満たす.

$$\mathrm{E}\left[u_{-\alpha}\left(\frac{q_\xi^\alpha(x|x)}{f^\alpha(x|x;h)}\right) - \mathrm{D}_\alpha\big(q_\xi^\alpha(y|x), f^\alpha(y|x;h)\big) \;\middle|\; h(\xi)\right] = 0. \tag{4.5}$$

(4.5) から, リスク最小問題 (4.2) の最小値が $u_{-\alpha}\big(c_x(h)\big)$ であることが分かる. これによって汎関数 $\psi_x(h)$ を次のように定義する.

$$\psi_x(h) := -u_{-\alpha}\big(c_x(h)\big). \tag{4.6}$$

Theorem 4.2. $s(\xi) = t_x(\xi;h)$ のときに限り, 制約付きの最小問題

$$\begin{aligned} &\min_{q(y|x)} u_\alpha\big(q(x|x)\big) \\ &\text{subject to } -\mathrm{D}_\alpha\big(q_\xi^\alpha(y|x), q(y|x)\big) + u_\alpha\big(q(x|x)\big) = s(\xi) \end{aligned}$$

は, リスク最小問題 (4.2) と同一の解 $f^\alpha(y|x;h)$ をもつ.

$u_\alpha(r)$ の単調減少性から, Bayesian log-likelihood $\log q(x|x)$ の制約付き最大問題とも等

価である．これは $\alpha=-1$ のときにも正しい．しかし，$\alpha=+1$ のときだけは特別扱いする必要がある．

等式 (4.5) に現れる量

$$u_{-\alpha}\left(\frac{q_\xi^\alpha(x|x)}{f^\alpha(x|x;h)}\right)$$

を最大化することによって canonical weight $h_x^\dagger(\xi)$ を定義する．$u_{-\alpha}(1/r)$ は r について単調増加なので，結局，次式によって定義される．

$$\delta_G \log f^\alpha\big(x|x;h_x^\dagger;h-h_x^\dagger\big)=0 \quad \text{for any } h. \tag{4.7}$$

Theorem 4.2 の後のコメントと逆に，この場合は $\alpha=+1$ のときにも正しいが，$\alpha=-1$ のときだけは特別扱いする必要がある．

Theorem 4.3. (4.7) によって定義される $h_x^\dagger$ を用いた予測分布 $f^\alpha(y|x;h_x^\dagger)$ は次の等式を満たす．

$$u_{-\alpha}\left(\frac{q_\xi^\alpha(x|x)}{f^\alpha(x|x;h_x^\dagger)}\right)=\mathrm{D}_\alpha\big(q_\xi^\alpha(y|x),\, f^\alpha(y|x;h_x^\dagger)\big) \quad \text{for any } \xi. \tag{4.8}$$

‘トップ’が最適予測分布であり，‘ビリ’が Theorem 4.3 の予測分布であるような予測分布のクラスを導くことができる．$h_x^*(\xi)=\lambda(\xi|x)$ とおく．

Theorem 4.4. 次の等式を満たす予測分布のクラスを $\mathcal{Q}^\alpha$ と書く．

$$\mathrm{E}\left[u_{-\alpha}\left(\frac{q_\xi^\alpha(x|x)}{f^\alpha(x|x;h)}\right)-\mathrm{D}\big(q_\xi^\alpha(y|x),\, f^\alpha\big(y|x;h\big)\big)\;\middle|\;\lambda(\xi|x)m(x)\right]=0.$$

また，(4.7) によって定義される $h_x^\dagger$ が実際に Bayesian log-likelihood $\log f^\alpha(x|x;h)$ を最大化するとする．このとき，$\mathcal{Q}^\alpha$ の中で，$f^\alpha\big(y|x;h_x^*\big)$ は最良であり，$f^\alpha\big(y|x;h_x^\dagger\big)$ は最悪である．

(4.6) で定義された $\psi_x(h)$ を h について‘最大化’することにより，事後リスクが prior

averaging density によらず一定となるような予測分布を導くことができる.

Theorem 4.5. Canonical weight $h_x^{\mathrm{c}}(\xi)$ を次式によって定義する.

$$\delta_G \psi_x(h_x^{\mathrm{c}}; h - h_x^{\mathrm{c}}) = 0 \quad \text{for any } h.$$

このとき, 予測分布 $f^{\alpha}(y|x; h_x^{\mathrm{c}})$ は次の等式をみたす.

$$\mathrm{D}\big(q_{\xi}^{\alpha}(y|x),\, f^{\alpha}(y|x; h_x^{\mathrm{c}})\big) = -\psi_x(h_x^{\mathrm{c}}) \quad \text{for any } \xi.$$

REFERENCES

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. Pages 267-281 in B.N. Petrov and F. Csaki (editors) *Second International Symposium on Information Theory.* Akademiai Kiado, Budapest.

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika*, **62**, 547-554.

Amari, S-I. and Nagaoka, H. (2000). *Methods of Information Geometry.* American Mathematical Society, Load Island.

Corcuera, J.M. and Giummole, F. (1999). A generalized Bayes rule for prediction. *Scandinavian Journal of Statistics*, **26**, 265-279

Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, **14**, 382-417.

Kullback, S. (1959). *Information Theory and Statistics.* Wiley, New York.

Ohnishi, T. and Yanagimoto, T. (2013). Twofold structure of duality in Bayesian model averaging. *Journal of the Japan Statistical Society, to appear.*

Stein, C.M. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, **9**, 1135-1151.

Yanagimoto, T. and Ohnishi, T., (2009). Bayesian prediction of a density function

in terms of e-mixture. *Journal of Statistical Planning and Inference*, **139**, 3064-3075.

Yanagimoto, T. and Ohnishi, T., (2011). Saddlepoint condition on a predictor to reconfirm the need for the assumption of a prior distribution. *Journal of Statistical Planning and Inference*, **141**, 1990-2000.